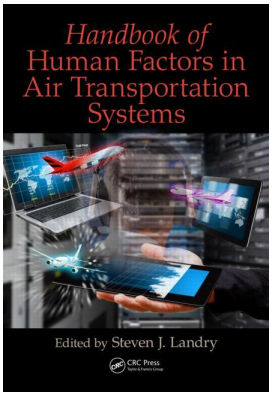

Handbook of Human Factors in Air Transportation Systems

Steven J. Landry

Data Sources and Research Tools for Human Factors

Publication details
https://test.routledgehandbooks.com/doi/10.1201/9781315116549-11
Mark Wiggins, Shayne Loft, Johanna Westbrook
Published online on: 15 Nov 2017

11 Data Sources and Research Tools for Human Factors

Mark Wiggins, Shayne Loft, and Johanna Westbrook

CONTENTS

Data Collection and Management in Aviation 239
Measures of Central Tendency 242
Distributions of Data 245
The Importance of Continuous Data Acquisition and Analysis 246
Organizational Trend Analyses 248
Data in Different Organizational Contexts 250
Acquiring and Using Data Cautiously 252
The Value of Qualitative Data 253
The Importance of Outliers 254
References 256

To remain competitive in a commercial environment, constantly innovating and searching for efficiencies, sources of data are required that reduce the uncertainty associated with decisions and guide a path toward improvements in performance. The business world is undergoing a revolution driven by the use of data and analytics to guide decision-making. Gone are the days where senior leaders within organizations could decide simply on the basis of intuition to invest in new technologies or a new route structure. In the contemporary industrial environment, regulators, consumers, and investors alike demand evidence-based and, importantly, defensible decisions from organizational leaders.

Defensible decisions, such as those in all walks of life, are dependent upon the availability of meaningful, relevant, and timely sources of data. The difficulty lies in the realization that both the sources and the amounts of data that are now available to aviation managers are almost endless, with aviation data analysts and managers facing a difficult choice in discerning reliable from unreliable data, meaningful from irrelevant data, and valid from invalid data.

Skilled aviation managers, similar to skilled military commanders, are those who have the capacity to quickly and effectively identify the key sources of information to which they should attend and at which time during the business cycle. This chapter explores the types and sources of data and methods of acquisition and analysis that are appropriate for the different conditions that aviation businesses might face now and into the future as they attempt to improve major aspects of their business, from using data to improve efficiency and safety, to improving customer retention.

DATA COLLECTION AND MANAGEMENT IN AVIATION

With 35 million flight departures per year, data are critically important for any planning decision made by aviation management. According to forecasts made by the International Civil Aviation Organization, passenger and freight air traffic will double by 2030 (ICAO, 2013). In the aviation industry, data are available on an almost continuous basis from a number of different sources, including aircraft sensors, air traffic control, passengers, customer search/booking data, and baggage screening. These are data at the micro level and provide the basis for monitoring various

components of the ongoing performance of a system. Changes in performance reflect responses to day-to-day or minute-to-minute demands that can be addressed quickly where necessary.

At the macro level, a complex organization might also rely on monthly or annual reports, such as profit and loss statements (Tretheway & Marhvida, 2014). These data are useful insofar they provide a cumulative assessment of the performance of an organization over an extended period and take into account the variability of performance across a period (Sull, 1999).

This notion of using macro data as a barometer of organizational performance is also evident in the safety context where, at its coarsest level, the frequency of incidents and in some cases, accidents over an extended period, is used to assess the performance of an organization in achieving its safety-related goals (Madsen, 2013; Travaglia, Nugus, Greenfield, Westbrook, & Braithwaite, 2011). Similar to profit and loss statements, these data take into account variability over an extended period. This is particularly important and useful in high reliability organizations in which the frequency of incidents and accidents may be relatively low, and/or the incidence is intermittent.

Just as these macro-level data offer useful indicators of performance, they are inevitably outcome-driven and suffer a potential lag between an occurrence and the reporting of that occurrence against a broader dataset (Lindberg, Hansson, & Rollenhagen, 2010). In the absence of other data, the result tends to be reactive and importantly, implemented sometime after an event or series of events have actually occurred. Therefore, the ideal approach is one that involves the collection of data on a cumulative basis, but at a level that is both meaningful and does not draw too heavily on organizational resources.

The importance of cumulative data, rather than a reactive approach to every case that occurs lies in both the efficiency and the effectiveness of the response (Jones, Kirschsteiger, & Bjerke, 1999; Lenne, Salmon, Liu, & Trotter, 2012). If an organization was to respond, in isolation, to every event that occurred, the demands on the organization would increase significantly, to a point where the demands may outstrip the resources available. The advantage of a cumulative assessment over a period of events or time is the opportunity to identify patterns of events, so that interventions impact performance beyond a single occurrence (Edkins, 1998).

The disadvantage associated with sole reliance on cumulative data is the potential delay in initiating a response until such time as a pattern of events emerges. To provide an example, consider a situation in which a single case of a disease with serious consequences is identified in a major population center. For the authorities not to respond immediately would be tantamount to negligence. However, a response that is disproportionate to the threat posed would also be criticized on the basis that valuable resources are being wasted or worse, drawn away from the management of other diseases that then reach epidemic proportions for want of resources. This is all the more important in commercial airline operations, as resources are inevitably constrained, and the costs of interventions can be significant.

This balance between the resources available and potential consequences is an important factor in determining the types of data acquired, the rate at which these data are acquired, and the nature of the response. This process of risk assessment to data acquisition and management is a necessary requirement in both commercial and noncommercial operations but should be approached with a clear understanding of both the severity and the likelihood of an adverse event.

In the case of severe consequences that occur infrequently, the difficulty lies in sourcing data on a sufficiently frequent basis that enables the identification of changes that signal a deterioration in performance. However, this approach presupposes that the precursors to severe events are understood, whether it is a minor conflict in the Middle East that escalates and causes a spike in oil prices, or a low pressure weather system over the Pacific that forms into a typhoon that impacts Hong Kong. In practice, these precursors are often only identified in the case of a system failure, such as an aircraft accident or incident.

An aircraft accident is rarely an isolated event. It is often the outcome of a process of less obvious failures or events that, had they been identified and rectified, might have prevented the accident (Goh, Love, Brown, & Spickett, 2012). These precursors or indicators are precisely the features that data acquisition systems such as Line Operations Safety Audit (LOSA) are intended to identify (Table 11.1).

TABLE 11.1
Summary of the Advantages and Disadvantages of the Various Approaches to Data Acquisition in Organizational Contexts

Approach	Sensitivity Management	Data Cost	Relative	Reliability
Continuous	High	High	High	Low
Cumulative	Moderate	Moderate	Moderate	Moderate
Summary	Low	Low	Low	High

LOSA is an audit process, much like a financial audit. Trained observers located in supernumerary seats on the flight deck record the behavior of pilots in response to the threats, errors, and undesirable states that occur during *normal* aircraft flights (Goodheart & Smith, 2014; Ma et al., 2011). Importantly, data are collected at every stage of the operation so that there is an opportunity to capture issues that might otherwise be overlooked using a more targeted approach. These deidentified data can be collapsed across routes, types of aircraft, or airlines. In addition to specific incidents that are identified as part of this process, comparative analyses are possible both within and across operations (Thomas & Petrilli, 2006).

In addition to more systematic strategies for data collection such as LOSA, it is now the case in most high-technology environments that data are collected automatically from a range of sources. In the context of human performance-related system failure, the masses of data that are acquired, referred to as Big Data, offer an opportunity to identify minor variations in behavior that might subsequently be associated with system failure. However, like systematic strategies for data acquisition, the utility of Big Data lies in the identification of the precursors to events. These precursors are what Walker and Strathie (2016) refer to as *leading indicators.*

In the context of human performance-related system failures, leading indicators are associated with psychological precursors of errors, so that a change in the leading indicator, in effect, becomes a barometer of the risk of a system failure. A practical example of a leading indicator might include changes in an operator's response latency to a series of warnings or alarms. An exceedingly rapid response might be indicative of a bias towards responding to the alarm, perhaps due to the fact that the alarm is always triggered at a particular point in the operational cycle. In the aviation context, the *minimums* warning on final approach to land is an example of an alarm that might be expected to be heard on the flight deck. A response that is too rapid might reveal a lack of conscious consideration of the event, whereas a response that is too slow might reveal a lack of preparation for the landing.

For aviation safety, the usefulness of LOSA and similar data collection systems lies in the capacity to identify *changes* in behavior as potential precursors to system failure. This allows the development and implementation of design, procedural, and/or personnel interventions that are targeted towards a specific issue that poses a threat to operational safety, thereby ensuring the efficient use of limited resources. However, it also benefits organizations in other ways, including the identification of inefficient procedures that increase the risk of delays or otherwise impact adversely the performance of the organization or system (Helmreich, 2000), and the identification of examples of *superior* pilot performance that can provide model behavior for application during training.

In addition to the identification of changes in performance over successful periods of data collection, the outcomes need to be compared against a benchmark. Benchmarks offer an objective standard of performance and can be prescribed by a regulatory agency as a minimum standard, or can be established by summarizing the normal or typical performance of other, similar organizations.

Performance norms are in common use in psychometric assessment and testing, in which it is difficult to prescribe appropriate levels of performance. The value of norms lies in establishing reasonable comparisons or expectations for performance. For example, it would be unreasonable to expect, during the initial stages of training, that a practitioner would demonstrate levels of

performance that are comparable to practitioners with much greater levels of experience. In the case of safety-related activities, no incidents or accidents is the ideal. However, in practice, the frequency of errors or incidents is highly variable, and the value of normed data lies in establishing whether a particular series of occurrences is a symptom of an underlying trend or is *normal* or typical (and then managed) given the context within which they occur.

MEASURES OF CENTRAL TENDENCY

Establishing benchmarks begins by establishing measures of central tendency that constitute aggregated representations of a dataset. The most common of these measures of central tendency is the *mean*, which is calculated as the sum of the data points, divided by the number of data points in a dataset. To establish measures of central tendency, it is important first to determine whether the dataset constitutes information from a complete population, such as all of the operators within an organization, or a sample of a larger population.

A sample is used when access to data from the entire population is difficult to obtain. This is not normally the case within organizations, although there are many examples in which large multinational organizations will sample a small group of employees and extrapolate these results to the organization more broadly.

It is this process of extrapolation that means that, inevitably, there is a degree of error. For example, it may be the case that the sample selected was drawn from employees within one particular sector or geographic location that does not necessarily represent the views or capabilities of employees in other locations. Statistically, this possibility can be taken into account by using different statistical analyses.

In the case of a population, the mean is calculated as

$$\mu = \frac{\Sigma(a^1 \cdots a^n)}{n}$$

where:

μ represents the mean
a represents the individual values
n represents the number of values in the dataset

In the case of a sample of a population, the mean is calculated as

$$\overline{X} = \frac{\Sigma(a^1 \cdots a^n)}{n-1}$$

where:

$\overline{X}$ represents the mean
a represents the individual values
n represents the number of values in the dataset

Although the mean is certainly the most common measure of central tendency, it needs to be calculated with caution as it is particularly susceptible to changes when the data are not distributed *normally*. In examining any phenomena for a period of time, behavior will be distributed across a normal or Gaussian distribution (essentially a unimodal, roughly symmetrical, bell-shaped curve) (see Figure 11.1). The nature of this distribution is best described in terms of human performance in which performance occasionally will be high or low but will generally cluster around the *mean*. The occasional forays into exceptionally high performance are reflections of statistical variability

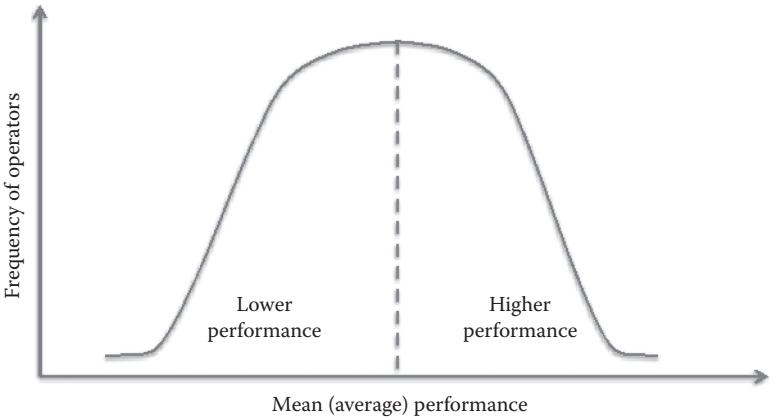

FIGURE 11.1 An example of a normal or Gaussian distribution in which the frequency of operators displaying higher levels of performance broadly matches the frequency of operators displaying lower levels of performance.

that occur for a variety of reasons. Therefore, they are not necessarily a reflection of normal performance. This is why the collection of data from a number of sources is so important. Any single data point may reflect performance that is the exception, rather than the rule.

To illustrate, an airline might be required to calculate the mean frequency of incidents to report to a regulatory authority. Incidents are calculated on a quarterly basis so that, over a 12-month period, the mean number of incidents is calculated as the sum of the total number of incidents, divided by the four quarters. This yields a mean frequency of incidents per quarter.

In a typical year, the number of incidents might be recorded as

Quarter 1–12 Incidents
Quarter 2–9 Incidents
Quarter 3–14 Incidents
Quarter 4–11 Incidents

Substituting,

$$\mu = \frac{\Sigma(12,9,14,11)}{4}$$

$$\mu = \frac{\Sigma(46)}{4}$$

$$\mu = 11.5$$

In this case, the airline experiences a mean 11.5 incidents per quarter.

However, if, during one quarter, the frequency of incidents was to increase markedly, it would influence significantly, the final result. For example, if the number of incidents was distributed as

Quarter 1–12 Incidents
Quarter 2–9 Incidents
Quarter 3–14 Incidents
Quarter 4–27 Incidents

Then, substituting*,

$$\mu = \frac{\Sigma(12,9,14,27)}{4}$$

$$\mu = \frac{\Sigma(62)}{4}$$

$$\mu = 15.5$$

This result is now much poorer than the previous 11.5. However, the majority of the results in the sample remain identical. The "27" incidents in the case of quarter 4 is referred to as an outlier, and it has affected significantly, the overall assessment of performance. Therefore, the overall result needs to be interpreted with some caution.

The main difficulty in using the mean as a measure of central tendency is the failure to account for outliers that may have skewed a distribution (see Figures 11.2 and 11.3). Although there may be very good reasons why an outlier may have occurred (such as the introduction of new aircraft or a period of poor weather conditions), it does not necessarily represent the pattern of the remaining data points. An alternative approach that better represents the dataset in this case is the median or *middle value*.

To establish the median, the values are ranked in order from least to greatest, and the middle value is taken as the measure of central tendency. In the case of the airline, there are four data points, so the middle value is calculated as the point half-way between the second and the third values (13). Where there is an odd number of items in a dataset, the middle value can be taken as the median.

A third measure of central tendency, although rarely used, is the mode. The mode is the most frequently occurring value in a dataset and is actually useful in establishing whether a ceiling or floor effect was evident. A ceiling effect occurs when a significant proportion of the values are recorded at a maximal value. For example, in an examination, the majority of students might

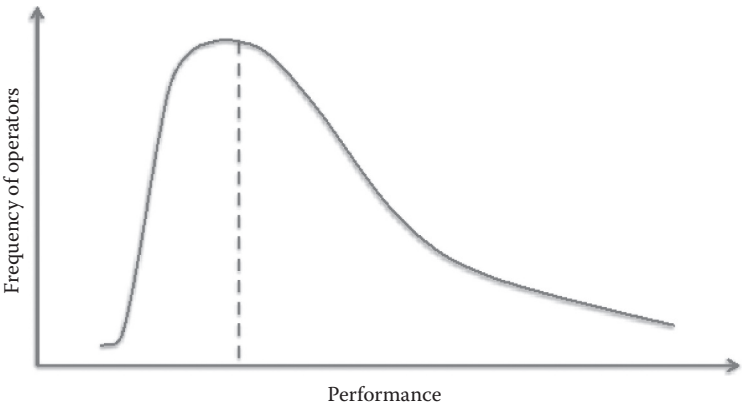

FIGURE 11.2 An example of a positively skewed distribution in which the frequency of operators is distributed across levels of performance. In this case, it is evident that, while the performance of the majority of operators clusters about the mean (signified by the dotted line), a small number of operators display exceptionally high levels of performance.

* Note that the population mean is used in this case since the data summarize the performance of the organization as a whole.

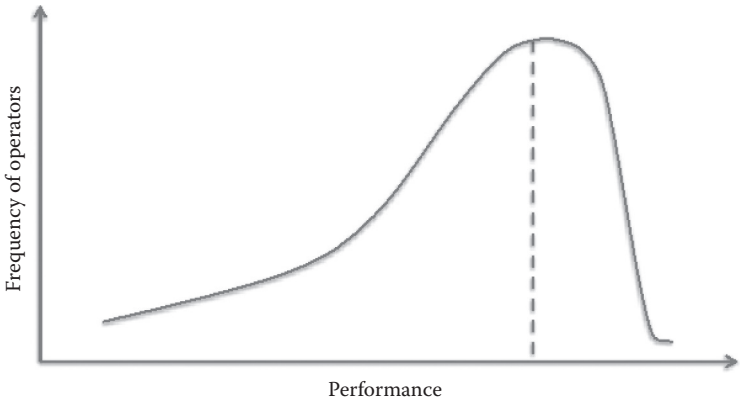

FIGURE 11.3 An example of a negatively skewed distribution in which the frequency of operators is distributed across levels of performance. In this case, it is evident that, although the performance of the majority of operators clusters about the mean (signified by the dotted line), a small number of operators display exceptionally low levels of performance.

achieve 100%. Conversely, a floor effect occurs where a significant proportion of data points are recorded at a minimal value. In either case, it reflects a systematic issue either in the way that the data were collected or within the examination itself.

Where an organization is safety conscious and is governed by highly practiced policies and procedures, a relatively high frequency of periods of no incidents (floor effect) might be expected. If, at the same time, the organization is operating as efficiently as is practicable, the frequency of on-time departures might be maximized (ceiling effect). Although these results are difficult to interpret from the perspective of the mean or median, they nevertheless represent very useful outcomes in establishing the effectiveness of policies, procedures, and practices.

DISTRIBUTIONS OF DATA

Although measures of central tendency are useful in summarizing the performance of the majority of values within a dataset, it is possible that, by focusing on these measures, problems with other aspects of the performance of an organization may be overlooked. For example, although a mean might reveal a high level of overall performance in the context of on-time performance, it may also belie the fact that there remain a number of data points that fall well below the mean and, in fact, these are likely to represent the key targets for intervention.

Falling below the mean is not necessarily indicative of an outlier. A distribution about the mean is to be expected in any dataset as values are naturally likely to fall about the mean. It is the breadth of this distribution that is of greatest interest in practice as it may reflect a lack of consistency within the cohort. For example, in the previous example, the frequency of incidents ranged from 9 incidents in Quarter 2 to 27 incidents in Quarter 4.

From an organizational perspective, the first task in determining whether there is an opportunity for intervention lies in calculating the nature of the distribution. Where the mean is used as a measure of central tendency, the standard deviation is used to represent the distribution. If the median was used, then the range of scores, from greatest to least, represents the distribution of values.

The standard deviation constitutes the mean squared differences about the mean and is calculated, for a population, as

$$\sigma = \sqrt{\frac{\Sigma\left(X - \bar{X}\right)^2}{N}}$$

where:

σ is the population standard deviation
X constitutes the individual values in a dataset
$\bar{X}$ represents the mean
N is the number of items in the dataset

Using the following dataset,

Quarter 1–12 Incidents
Quarter 2–9 Incidents
Quarter 3–14 Incidents
Quarter 4–11 Incidents

the standard deviation is calculated, as

$$\sigma = \sqrt{\frac{13}{4}}$$

$$\sigma = \sqrt{3.25}$$

$$\sigma = 1.80$$

Therefore, in the case of this dataset, the mean deviation about the mean is 1.80. This result can now be compared with data from other periods to determine whether there are any differences in the distribution of scores. A change in the distribution might be indicative of a change in the performance during one or more quarters. This may not have affected the mean, as a loss of performance for one Quarter might be counterbalanced by an improvement in performance during another quarter. However, by examining the distribution, it becomes clear that a change may have occurred that requires further analysis.

Where the data are not normally distributed and the median is selected as the measure of central tendency, the standard deviation is no longer an appropriate measure of variability. In this case, the range needs to be calculated. The range is simply the lowest score subtracted from the highest score. This gives an indication of the breadth of scores.

The standard deviation and variance are only of practical use where the outcomes can be compared against other datasets. For example, if data are collected on an annual basis, it is possible to compare the distributions from one year to another and establish whether any changes have occurred, either in terms of central tendency or in terms of the distribution of values.

From an organizational perspective, differences in the distribution of values, even in the absence of differences in measures of central tendency, constitute an indicator of changes in organizational performance. Specifically, an increase in the standard deviation or range is a reflection of increased variability so that, although there are periods of outstanding performance, there are also periods of poorer performance. As a consequence, examining both measures of central tendency and the distribution can provide the impetus for examinations of the causes of poor performance and the features that enable successful performance. By drawing on the factors that contribute to success, it becomes possible to target areas of poor performance, thereby improving measures of central tendency and reducing the standard deviation.

THE IMPORTANCE OF CONTINUOUS DATA ACQUISITION AND ANALYSIS

For most assessments of organizational performance, the value of measures of central tendency can only be realized once the values are compared with a normed dataset. These are datasets (a) from the same group but at different periods or (b) from different, but comparable groups. It is only by

undertaking these types of comparisons that changes or differences within and/or between organizations can be identified and their impact interpreted. However, this comparative process needs to be undertaken with a level of caution to ensure that the assessments are valid.

Clearly, comparative groups need to be considered carefully to ensure that there are no systematic differences that might render the assessment invalid. For example, if the comparative group operates different aircraft, operates to different destinations, or differs substantially in size, then some comparisons, such as the mean number of days lost to workplace injury, might be unreasonable. Even assessments of safety-related performance might lack validity where one organization operates aircraft to and from high-capacity airports where another operates aircraft to uncontrolled airports. In this case, the nature of the hazards will differ and although both organizations will comply with regulatory requirements, the expectations pertaining to the nature and frequency of incidents need to be considered.

Norms overcome the problem of valid comparisons by providing a stratified benchmark so that datasets are comparable. In the case of organizations, this stratification might result in levels of performance that are assessed and are then distributed along one or more dimensions. The number of employees or the number of aircraft operated might constitute two of these dimensions. By matching the characteristics of an organization to the appropriate levels of these dimensions, the analysis becomes comparable.

The main difficulty associated with the establishment of norms is that they require multiple sources of data for each level of a dimension so that means can be calculated that reflect the breadth of a distribution. This process can take time and considerable resources to accomplish, ensuring that the data are collected using the same data collection tools and techniques. LOSA offers an ideal opportunity in this regard, as the process of data collection is standardized, and data are collected from multiple organizations across the world.

Calculating performance against a norm begins by taking into account the relevant dimensions, such as the size of the airline, and then identifying the mean industry performance for airlines of this size. The z score indicates, for a particular organizational value, the number of standard deviations it sits above or below the industry mean. It is calculated as

$$z = \frac{X - \overline{X}}{\text{SD}}$$

where:

X represents the organizational value
$\overline{X}$ represents the industry mean
SD represents the standard deviation

Using an organizational value of 11.5 incidents per year, where the mean number of incidents for organizations of a similar size is 13.6, with a standard deviation 2.3, the values can be substituted as

$$z = \frac{11.5 - 13.6}{2.3}$$

$$z = \frac{-2.1}{2.3}$$

$$z = -0.91$$

indicating that the outcome for the organization is almost one full standard deviation below the industry mean. This is a strong result, although it represents performance over a single year.

The question remains as to whether this result is sustained or even improved over time. It requires the collection and analysis of data on a systematic basis by both the organization and the industry more broadly.

An allied issue that needs to be addressed in relation to the collection and analysis of data concerns the rate at which these data are optimally acquired. Inevitably, the implementation of audit systems is costly and needs to be undertaken judiciously, but with sufficient frequency to enable comparative analyses. The value in these types of analyses is both in identifying specific instances and trends that, over time, might constitute a catalyst or precursor to system failure.

ORGANIZATIONAL TREND ANALYSES

Trend analyses are a key component in identifying and responding to conditions where there is a lag between an input and an output. For example, the effects of a restructure at senior levels within an organization may take sometime to translate into changes in organizational performance at the operational level. Therefore, establishing the effects requires the acquisition of data at a range of intervals following the introduction of the change.

Trend analyses are also evident in the context of aviation safety in which the frequency of accidents at a specific point in time is likely to be too few to draw any meaningful conclusions. It is only in the accumulation of data that a pattern becomes evident (O'Hare, Wiggins, Batt, & Morrison, 1994; Wiegmann & Shappell, 2001; Wiggins, Hunter, O'Hare, & Martinussen, 2012). In this case, trend-type historical data are used to predict future outcomes. This pattern might be as simple as a greater frequency of one type of event relative to other events. At a more complex level, it might be the case that a particular event occurs more frequently, but only at a particular point in time. To isolate the issue further, it may become clearer that an event only occurs at a particular time when it is preceded by another factor. Trend analysis has been used to quantify the contribution of the frequency and severity of maintenance errors to aircraft accidents and incidents. In the case of Marais and Robichaud (2012), maintenance-related aircraft accidents were 6.5 times more likely to be fatal than accidents in general and, when fatalities did occur, maintenance accidents resulted in 3.6 times more fatalities on average than aircraft accidents in general.

A relative newcomer to predicting the impact of aviation historical risk factors involves Bayesian Belief Networks (Brooker, 2011). The networks use aviation experts to estimate the probabilities of events. The probabilities are conditional; The chance of something happening given that something else has happened. Probabilities are then combined to model the probabilistic behavior of the system. Bayesian models are well established (for a review see Kruschke, 2015), but part of the challenge is that it can be difficult for aviation experts to estimate rare event conditional probabilities accurately (i.e., Bayesian priors; Kruschke, 2015). Nonetheless, there have been several demonstrations of how Bayesian models can possess a satisfactory range of accuracy in predicting aviation risk (Feng, Sahin, & Karson, 2009; Wang & Gao, 2013). Wang and Gao (2013) used Bayes to quantify the incremental risk to safety caused by flight delays in Chinese airspace, and Feng et al. used Bayes to evaluate the tradeoff between system risk and system cost with different baggage screening systems.

Although Bayes' estimates have been applied successful in a number of contexts, the utility of the approach tends to diminish with increases in complexity or where the specific influence of a variable is unknown. An example of the complexity of these relationships can be drawn from investigation into a series of BEA Comet crashes that occurred in 1954 (Simons, 2013). The Comet was the first commercial jet airliner in service, and much was riding on its success. When, inexplicably, two aircraft broke up over the Mediterranean Sea, it was critical that the cause, or causes, of the failure were established.

The problem for the investigators of the Comet crashes was a problem encountered in subsequent aircraft investigations; the lack of data that had been collected under controlled conditions. Unless specific precursors to the events could be recreated and various combinations compared, it would be

unlikely that the specific causes could be identified. In the event, investigators and researchers at the Farnborough Research Laboratories spent many weeks subjecting the mock-up of a Comet Fuselage to different levels of simulated air pressure and importantly, the repeated increase and decrease in pressure that was to constitute the catalyst for fatigue cracks that emerged in and around the passenger windows (Simons, 2013).

The passenger windows in the Comet were a square shape, rather than the oval shape that was used in later, high altitude aircraft. This design, together with the method of manufacture, created the preexisting conditions for fatigue-related stressors. The stress imposed by constant depressurization and repressurization inevitably resulted in the fatigue-related fractures that, in turn, caused explosive decompression and the loss of the aircraft (Withey, 1997). However, it was only due to the methodical collection and analysis of data under a variety of controlled conditions that resulted in the mystery being solved.

The systematic approach to understanding the causes of the so-called Comet crashes is the same approach that researchers use to understand other complex phenomena. It relies on samples collected over time and/or under different conditions. This sampling approach accounts for differences in natural or stochastic variability that influences responses from time to time (Wiggins & Stevens, 1999).

The collection of data across a random sample is the next phase in establishing valid assessments of performance. This is the advantage of tools such as LOSA insofar as data are collected across a broad and random sample of operations and crews. The intention is to establish the level of performance in general, taking into account the performance of crews that are at the extremes (Ma et al., 2011).

Clearly, rare but severe consequences, such as the loss of an aircraft, demand an investment in sources of data that will enable the causes to be identified and addressed. However, even in this context, it is possible to use sample-based data to identify the causes of a failure and/or establish the prevalence of factors that will enable the identification of potential failures in other jurisdictions or amongst other carriers.

A useful example of the acquisition of these data can be drawn from the response of TransAsia Airways following the crash of an ATR-72 following an engine failure immediately after takeoff from Taipei International Airport. The airline undertook to evaluate the performance of its pilots following the accident (Ramzy, 2015). Although the airline probably should have been evaluating its pilots proactively, it does demonstrate the value of a sample-based approach to data acquisition following a major incident or accident.

Although the focus thus far has been directed toward the collection of safety-related data, airlines and aviation-related organizations have access to data well beyond safety and security. Airlines can collect and analyze data from a wide range of sources, including customer search and transactional (demographic data, online search behavior, preferences) data. As commercial concerns, the survival of airlines depends upon a reliable stream of revenue. One of the most important sources of revenue for commercial operations including airlines is daily sales. Sales data are available against daily or weekly costs (Simoudis, 1996). These data translate into the Load Factor that represents the average number of passengers per flight (Jenatabadi & Ismail, 2007). Normally expressed as a percentage or ratio, it allows an organization to quickly and reliably establish the cost effectiveness of a route or scheduled service.

A continuous data stream is a useful means by which changes in, and the performance of, an organization can be quickly and clearly established. The delay between the acquisition of the data and its receipt by analysts, including senior managers, is minimized. This, in turn, reduces the period between the acquisition of information and the implementation of a response.

The ability of organizations to quickly identify and respond to changes in the marketplace is a key precursor to sustained success (Hoogervorst, 2004). This agility is dependent upon the availability of accurate data, in the appropriate form, and at the appropriate time. Importantly, the appropriateness of data is possibly the most vexed in a technical environment where data pertaining to almost any variable are available on an almost continuous basis (Behn & Riley, 1999).

DATA IN DIFFERENT ORGANIZATIONAL CONTEXTS

One of the most significant shifts in air travel in recent years has been the transition from full-service to low-cost carriers. The latter has a lower cost-base, the savings from which can be passed to passengers in the form of lower ticket prices. With less reliance on infrastructure, these organizations are arguably much more agile in the marketplace and can quickly shift routes should there be changes in market demands. However, these carriers are also significantly dependent upon short turnaround times, so that there is little room for delays. Therefore, although the utilization of unproductive resources provides these carriers with a degree of flexibility, it also raises risks since low-cost carriers are especially dependent upon the availability of real-time data that would enable the anticipation of delays or system failures (Reddy & Reddy, 2002).

In the Boeing assembly plant in Seattle, every part that is used in the manufacture of an aircraft is electronically labeled so that, at any one time, it is possible to establish the rates at which particular parts are consumed, when specific parts are required, and the costs of the purchase of parts (Eriksson, 2011; Funk, 1995). Replacement parts are purchased in a just-in-time strategy that ensures their availability without the costs of retaining large quantities of stock. However, this approach is only as effective as are the processes and capabilities of the suppliers of that stock. In the case of Boeing, the materials used are constructed from very high-quality materials and are subjected to rigorous testing. Therefore, data pertaining to the consumption of materials needs to occur in sufficient time to enable manufacturers to undertake the necessary construction and testing, prior to delivery.

Although the consumption of components enables organizations such as Boeing and Airbus to enjoy efficiencies, it also provides a surrogate measure of the performance of different parts of the organization. In effect, the consumption of parts represents a barometer in reaching organizational performance targets which, in the case of companies such as Boeing, Airbus, and Embraer, is the construction of new aircraft. As there is an inevitable delay in the construction of aircraft, the consumption of parts might be employed as a means of identifying gaps and/or delays in the construction process.

Although incident reporting systems provide crucial information for many industries, understanding of the strengths and limitations of incident data is often poor. In the health sector, underreporting is recognized as a serious limitation (Ohrn, Elfstrom, Liedgren, & Rutberg, 2011; Stanhope, Crowley-Murphy, Vincent, O'Connor, & Taylor-Adams, 1999; Westbrook et al., 2015). The incidents reported often fail to represent either the frequency or the distribution of the types of errors that occur. For example, a study comparing medication errors identified during a detailed review of 3,291 patient records with incident reports showed that only 13 of 1,000 errors identified had been reported to the incident system (Westbrook et al., 2015). Evidence in patient records that staff had detected but not reported a serious error occurred at a rate of 219 of 1,000 errors. However, for 78% of the serious medication errors (with the potential to lead to significant patient harm) identified, there was no evidence that staff had in fact detected the errors. Thus, the reliability of the data within incident reporting systems is both heavily reliant upon the reporting culture, and on the vigilance and skills of the staff in error detection.

In some cases, the data from incident management systems can reveal more about an organization than simply the frequency or severity of errors. For example, a relatively high incident rate may actually reflect a positive safety culture and low risk compared to organizations with a relatively lower incident reporting rate. Evidence to support this proposition can be drawn from the medical context in which hospitals with higher incident report rates experience significantly fewer medication errors on audits than sites with lower incident reporting rates (Westbrook et al., 2015). A similar effect is evident in the aviation context where the frequency of incident reports is interpreted as a barometer of the safety health of an airline.

Despite the temptation to draw inferences on the basis of incident data, such analyses, between and across time, or organizations, must be undertaken with caution. The frequency of incident

reports pertaining to an event may not necessarily be associated with increased risk, and responding to small numbers of incidents by type over time is likely to reinforce perceptions that small changes in the frequency of reports are significant. Such interpretations may lead decision-makers to assign resources to address incident types that are easy to report, rather than those that pose the greatest threat to safety (Levinson, 2012).

Small numbers of incidents and the absence of denominators exacerbate the risk of misinterpretation of incident data (Shojania, 2008). Many industries experience serious incidents as relatively rare events. A good example is the case of drug maladministration in nuclear medicine (Larcos, Collins, Georgiou, & Westbrook, 2014). Such errors can have catastrophic effects and, as a result, mandatory reporting registries are in place in many countries. However, the incidence of these errors is estimated at only 6 per 100,000 nuclear medicine procedures (Larcos et al., 2014). Therefore, at an organizational level, incidents occur infrequently, and changes in the frequency of these events are very difficult to interpret.

An exciting new development in the study of risk is the work of Didier Sornette and colleagues on what they term *Dragon Kings* (Sornette & Quillon, 2012). These events differ from *Black Swans* (Taleb, 2007). The latter are extreme events that occur infrequently but still conform to the statistical distribution and are caused by the underlying processes of the scenarios under examination. So, in the case of accidents, hazards, or disasters, a Black Swan even though improbable (rare), is still an expectable outcome of the processes that generates more typical events. To give a specific example, earthquakes occur commonly along plate faults. Most are relatively small (even unnoticeable), but a largish earthquake that inflicts damage, although rare, is still to be expected eventually. However, because Black Swans result from the same underlying processes of typical events, they are essentially random and thus, individually essentially unpredictable. We know in the long run, these rare events will occur, but their actual occurrence is essentially random. Dragon Kings, however, are different. They are events that are too large or too impactful to be considered as occurring simply as an uncommon outcome of typical events. Instead, they are the result of run-away processes that cause a new order in the system to emerge.

In the language of complex systems or nonlinear dynamics, Dragon Kings are emergent phenomena that are indicative of a change in regime or the underlying dynamics of the system. These events are the result of a self-organizing dynamical transition which itself is due to run-away processes occurring in the feedback loops and interconnectedness of the system. Unlike Black Swans they are not, therefore, random extreme outcomes of the typical processes. Instead, they are new, emergent features or the result of a phase transition.

The one upside, if this is true, is that, unlike Black Swans that are individually unpredictable (random), Dragon Kings may have telltale cues indicating their emergence and there are available statistical tests to differentiate Dragon Kings from Black Swans (Janczura & Weron, 2012; Sachs, Yoder, Turcotte, Rundle, & Malamud, 2012). In simplistic terms, to detect an oncoming Dragon King, one needs to detect a phase transition in a dynamical system. This statistical approach has been employed in other settings to determine if a phase transitions occurs in individual skill development (Helton, 2011). New ways of doing a task, a new system phase, may be preceded by the cues indicative of phase transitions, namely an increase in performance variability.

As any system transitions from one state to another, a phase transition, there tends to be an increase in overall system variability. Mathematically, this is characterized as a shift in the system to a new attractor or influential point in dynamical space. This may be indicated by techniques that measure an increase in variability in a time series. The new attractor of the system causes system states nontypical of the previous attractor that is still influencing the system until the transition is complete. Tools such as running standard deviations or sample entropy could be employed to detect the increase in system variability prior to the complete phase transition. To accomplish this, however, relevant data need to be collected for a sufficiently long time series to be useful. The key to plausible prediction and thus, intervention, is rich data collection, storage, and ongoing analysis. These techniques are being employed in a wide variety of fields, from natural disasters to medicine, and may be useful to the aviation industry.

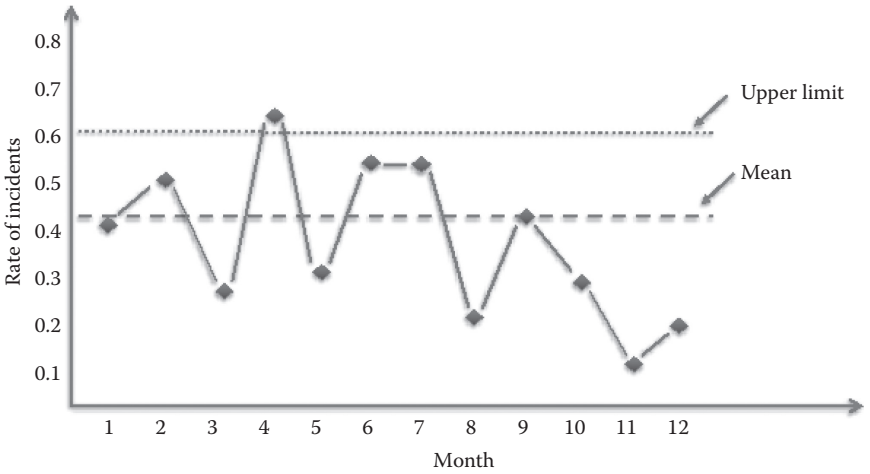

FIGURE 11.4 An example of a process control chart for the rate of incidents within an airline, distributed across a 12-month period and incorporating a designated upper limit that would trigger an intervention.

Statistical process control uses *control charts* to graphically monitor incidents to determine whether a process is in a state of statistical control and may be a useful approach to the monitoring of rare events (Benneyan, Lloyd, & Plsek, 2003; Larcos, Collins, Georgiou, & Westbrook, 2015; Montgomery, 2012). A control chart has a central line that represents the average or target value and an upper and lower line indicating the boundaries for the control limits. This approach recognizes that, for all processes or indicators, some variation is to be expected. The control limits are designed to detect when indicators move beyond natural, random variation and exceed the boundaries as a result of *special cause variation* (Laney, 2002). When a value falls outside the control limits, this signals that action may be required to bring processes back into control. Data patterns that may suggest special cause variation include, for example, one point outside three standard deviations of the mean or a run of at least six data points either increasing or decreasing. There are a range of different types of control charts that can be applied to data of different types. A U-chart is based on a Poisson distribution and presents counts as rates; a C-chart is also based on the Poisson distribution and presents the counts as numbers; and a P-chart is based on a binomial distribution and is used for percentages (Figure 11.4).

ACQUIRING AND USING DATA CAUTIOUSLY

As with all datasets, the information available can be used for a range of different purposes, depending on the nature of the organization and its goals. The difficulty occurs where data are used for a purpose for which they were not designed, or anticipated for, when collected. At the very least, in extrapolating a dataset across an organization, it is important to consider whether the data are representative of the broader population. If the data were sampled on single day, in a single geographic location, and/or amongst a particular group within the organization, it is possible that different responses might have been acquired, had the process of data collection occurred beyond a single day, and across the different geographic locations and assembly lines that might characterize the organization.

To extrapolate from a sample to a population is reasonable, so long as the sample is representative of the broader population to which the outcomes will be applied (Wiggins & Stevens, 1999). This requires some consideration of what are referred to as *extraneous variables*. These are factors that may systematically influence the responses that are elicited. As a case in point, evaluating people on a test of cognitive ability will yield different scores, depending upon the time of day at which the

test was administered (Revelle, Humphreys, Simon, & Gilliland, 1980). For the majority of people, testing late in the afternoon after a full day of work will result in mean scores that are lower than the same group tested earlier in the day.

Similar to fatigue, other extraneous variables can impact performance, including age, geographic location, experience, and the type of work undertaken. Therefore, in acquiring data from a representative sample across multiple parts of an organization, it is important to both ensure that the sampling is stratified, and that the impact of the stratification is evaluated prior to conclusions being drawn. Stratification is a process of sampling that involves first, the identification of those extraneous variables that are likely to differentially impact the outcomes. Thereafter, researchers deliberately source data that are representative of these different variables. Finally, comparisons are undertaken to ensure that the proportion of respondents or participants reflects the proportion of participants that comprise the organization. To put it simply, if women comprise 40% of the population of an organization and sex represents an extraneous variable, then females should comprise 40% of the sample.

Where it is not possible to stratify the process of data acquisition, perhaps because the demographic characteristics are unclear or it would undermine the process of data collection, it remains possible to assess the influence of various factors on responses and thereby determine the role of the extraneous variable in contributing to the overall outcome. For example, in assessing the performance of pilots' use of engine-out operations following takeoff, poor performance amongst a sample of pilots might be explained by a lack of recent training experience. Although it may not be possible to stratify the data collection process in this case, by collecting data about pilots' recent training experience, it becomes possible to account for these differences in performance statistically using methods such as Hierarchical Multiple Regression or Analysis of Covariance. These data analytic techniques enable a determination of the level of variability in the outcome variable (e.g., pilot engine-out performance) that is uniquely predicted by certain operator or environmental variables while controlling for other variables (e.g., training).

THE VALUE OF QUALITATIVE DATA

It is often difficult, prior to the collection of quantitative data, to identify explanatory variables and acquire data in a form that will enable the appropriate application of the aforementioned statistical control. Therefore, in addition to quantitative data, it is often useful to collect qualitative data that might explain some of the effects that are identified (Tariq, Georgiou, & Westbrook, 2013; Westbrook, Gosling, & Coiera, 2004). As an example, responses to a questionnaire concerning perceptions of safety within an airline might be influenced by both recent changes in procedures or by changes in management, and this possibility may only become evident following the completion of data collection. Therefore, there needs to be some *a priori* understanding of the context within which data are being collected.

In addition to the features of samples and the need to consider the impact of extraneous variables on the process of data acquisition, it is important to consider the nature of the data being acquired. For example, the acquisition of subjective data, such as attitudes or preferences, can mask behavior that might occur in practice. This is often the case where questions relate to a moral or ethical dilemma (Krumpal, 2013). How people respond to questions regarding safety may belie the fact that, under certain circumstances, safety behaviors do not necessarily reflect the attitudes that are espoused in response to questionnaires. Therefore, the ideal approach is one that draws together subjective responses with objective data. For example, an audit of a maintenance facility might highlight an increased incidence of defects that are undetected before an aircraft is returned to operations. The results of a questionnaire distributed subsequently may provide an explanation insofar as the results reveal a perception of under-resourcing and a lack of support from management.

Although cross-referencing data do not necessarily provide a complete explanation for an effect, at the very least, other forms of data can be used to either confirm or disconfirm an outcome. This is referred to as *triangulation*, and it can be used where there is a limited sample or an outcome is

of such importance (such as safety-related issues) that it requires further investigation (O'Connor, Buttrey, O'Dea, & Kennedy, 2011; Tariq, Georgiou, & Westbrook, 2013).

Although there was, for many years in the aviation industry, a propensity toward quantitative data, there has been a shift in recent years toward the collection of more qualitative data both as a means of exploring and explaining the quantitative data and in terms of identifying opportunities for new and potentially disruptive innovations (Wilke, Majumdar, & Ochieng, 2014). Due to the largely subjective nature of both the collection of qualitative data and its interpretation, drawing conclusions exclusively on the basis of qualitative information is fraught with danger, particularly if the data were drawn from a limited sample. However, there are some techniques that can be employed to assist with the interpretation of data and ensure its reliability and validity. The first of these strategies involves a clearly defined process of qualitative data categorization so that different analysts will reach the same or similar conclusions (Roth, 2015). This is important as it provides a degree of confidence to the effect that the interpretations have some basis in reality.

Categorizations can be determined prior to the collection of data or as part of an expectation based on anecdotal observations, previous research, and/or a particular theoretical perspective (Cassell & Symon, 2011). The analysis of air traffic controller self-reported errors conducted by Shorrock (2005, 2007) is a case in point. Shorrock interviewed UK air traffic controllers using an unstructured interview format that allowed controllers to free-recall and explore personal accounts of the kinds of errors they have made in the past, and any factors that have influenced their performance. Shorrock then used an error classification and analysis system called the technique for the retrospective and predictive analysis of cognitive error (TRACEr—Shorrock & Kirwan, 2002) to classify these data into various forms of human memory and perception errors.

However, whether categorizations of qualitative data, such as those documented by Shorrock (2005, 2007), can be proposed depends on the stage of the research at which data are being acquired. For example, in the case of levels of safety climate within a newly established airline, there may be few expectations as to the level of safety climate at a given point in time. Therefore, the process of data acquisition and interpretation is, inevitably, going to be driven by the nature of the data collected. However, as discussed previously, in the absence of a benchmark or expectation, the data do little to inform the organization as to whether or not the outcomes are satisfactory.

At an operational level, establishing a benchmark or expectation represents an anchor point around which future datasets can be compared. In fact, this is precisely the approach that is undertaken in the context of financial reporting, whereby annual profit or income is compared against performance during the preceding year. It is this capacity for a comparative assessment that enables prospective investors to establish the long-term performance of the organization. Similarly, in the context of safety, comparative data are necessary to draw meaning from numerical data (Oster, Strong, & Zorn, 2013).

Although the assessment of data against expectations or a hypothesis offers advantages, it is also possible that data will be interpreted in the context of these expectations and without due regard to the statistical properties of the dataset. For example, drawing inferences based on measures of central tendency, such as the mean (average), can be misleading, particularly if the data are non-normal. The distribution of these collected variables will not match the normal curve perfectly, but for many statistical analyses, referred to as *parametric analyses*, the distribution of the data needs to approximate a normal distribution. Where the data are non-normal, a parametric statistical technique is no longer representative of the dataset and a nonparametric technique needs to be employed or any outliers may need to be transformed or removed from the dataset.

THE IMPORTANCE OF OUTLIERS

By definition, an outlier is a data point that falls three or more standard deviations from the mean. They arise where there are systematic drivers influencing performance that do not impact the performance of other members of the dataset. For example, if pilots' performance during a standard

arrival procedure is being assessed, and a small proportion of the sample complete the assessment having just flown the approach, their performance may be systematically greater than pilots who have yet to fly the approach in practice.

In effect, outliers constitute a population distinct from the remainder of the sample. A systematic variation in demographic characteristics or the process of data collection has enabled the acquisition of data from a population distinct from the population under investigation. In some circumstances, it may be appropriate to simply remove extreme outliers from the sample, or to transform the dataset using a mathematical procedure (e.g., a square root transformation), that results in a sample distribution closer to a normal curve. However, it is important to recognize that outliers, although perhaps unrepresentative of a target population, could constitute a window into other populations that may be of interest.

In the context of aviation safety, employees performing particularly poorly or responses that are in marked contrast to the responses of other parts of an organization might be regarded as outliers given the responses from the broader sample. However, they also constitute an area of concern insofar as identifying risks to the organization. In the case of safety climate, outliers might reflect issues that need to be addressed, such as ineffective leadership, the lack of resources necessary to undertake tasks successfully, and/or a lack of appropriate training. In any of these cases, the outlying responses will have revealed an opportunity for intervention.

Outliers may also reveal something about the process of data acquisition and particularly whether the process of data acquisition was consistent across the sample. For example, differences in responses to a survey across different parts of an organization might simply reflect differences in the way in which a survey was distributed. In one case, the survey may have been distributed by a line manager, thereby encouraging a degree of compliance both in completing the survey and providing responses that might be regarded as socially or organizationally acceptable. In another case, the survey may have been distributed by the safety department, and this strategy might result in altogether different responses given that anonymity is assured.

As the identification of dependencies is the goal of most investigations, it is important to establish the extent to which the contribution to a global outcome is influenced by the contributions of different subgroups that comprise a broader cohort. By taking account of behavior at different levels within an organization, it becomes possible to establish the relative contribution of variables. For example, if the population of an airline comprises a series of subgroups, it might be determined that a particular combination of subgroups impacts differently, the overall responses of the cohort. This source of variance is important to establish as it provides a level of explanation that can be useful in determining where, for example, to invest training or other resources.

Referred to as multilevel modeling, or sometimes known as *random coefficient* or *mixed effect* models, this analytical process has been used widely in the biological and social sciences to analyze nested data structures (Hofmann, Griffin, & Gavin, 2000), and in the aviation context, it can be used to identify changes in performance over time whereby an intervention may affect differently a particular group with a cohort. The opportunity in this case is to establish whether interventions have been successful and, importantly, for whom within a cohort they were most successful or unsuccessful. This informs decisions, including how and to whom interventions might be targeted, thereby enabling the efficient and effective utilization of resources.

An example of the practical application of multilevel modeling in aviation can be drawn from an investigation of the workload management of Australian en route controllers (Neal et al., 2014). Managers of air traffic control systems proactively manage controller workload by redesigning airspace and procedures and adjusting rosters to ensure that anticipated task demands can be met. This requires the prediction of controller workload that, in effect, is a prototypical example of a multilevel problem as variability exists across different airspace sectors and between the capacity of different controllers (Loft, Sanderson, Neal, & Mooij, 2007).

Within sectors, air traffic controllers vary in their response to task demands as a consequence of individual differences. There is also a degree of variability over time, as controllers encounter

nonroutine events that fall outside the bounds expected under normal conditions. Finally, there are likely to be differences in the way that supervisors and fellow controllers make judgments of workload when observing the performance of their colleagues. Neal et al. (2014) developed a multilevel statistical model capable of predicting the range within which controller's subjective estimates of workload would be likely to fall with a certain probability (prediction interval) given the specific anticipated operational conditions (sector type, daily flight plans, personnel, available automation). Air traffic control management can use this predictive model to analyze historical and projected workflows to dynamically allocate resources to meet changes in demands.

One of the significant criticisms of evidence-based interventions is that the data on which the intervention is based were collected in single instance so that it becomes unclear whether the responses represent systemic issues or whether they are simply a reflection of the state of affairs at a particular point in time. This is important as it may misdirect both the need for an intervention and the scope of any intervention that may be applied.

The solution to the problem of one-off or cross-sectional forms of data collection lies in the acquisition of data over extended or multiple periods. This enables the assessment of intervening variables and ensures that the data collected in one instance are a robust and reliable reflection of the domain.

The assessment of intervening variables can be difficult to establish and requires the application of specialist statistical techniques. In essence, the goal is to determine whether the relationships between two variables are better explained by considering a third or more variables.

REFERENCES

Behn, B. K., & Riley, R. A. (1999). Using nonfinancial information to predict financial performance: The case of the US airline industry. *Journal of Accounting, Auditing & Finance, 14*(1), 29–56.

Benneyan, J. C., Llyod, R. C., & Plsek, P. E. (2003). Statistical control process as a tool for research and healthcare improvement. *Quality & Safety in Health Care, 12*, 458–464.

Brooker, P. (2011). Experts, Bayesian belief networks, rare events, and aviation risk estimates. *Safety Science, 49*, 1142–1155.

Cassell, C., & Symon, G. (2011). Assessing "good" qualitative research in the work psychology field: A narrative analysis. *Journal of Occupational and Organizational Psychology, 84*(4), 633–650.

Edkins, G. D. (1998). The INDICATE safety program: Evaluation of a method to proactively improve airline safety performance. *Safety Science, 30*(3), 275–295.

Eriksson, S. (2011). Globalisation and changes of aircraft manufacturing production/supply-chains—the case of China. *International Journal of Logistics Economics and Globalisation, 3*(1), 70–83.

Feng, Q. M., Shain, H., & Karson, M. J. (2009). Bayesian analysis models for aviation baggage screening. *IIE Transactions, 41*, 995–1006.

Funk, J. L. (1995). Just-in-time manufacturing and logistical complexity: A contingency model. *International Journal of Operations & Production Management, 15*(5), 60–71.

Goh, Y. M., Love, P. E., Brown, H., & Spickett, J. (2012). Organizational accidents: A systemic model of production versus protection. *Journal of Management Studies, 49*(1), 52–76.

Goodheart, B. J., & Smith, M. O. (2014). Measurable outcomes of safety culture in aviation—a meta-analytic review. *International Journal of Aviation, Aeronautics, and Aerospace, 1*(4), 1.

Helmreich, R. L. (2000). On error management: Lessons from aviation. *BMJ: British Medical Journal, 320*(7237), 781.

Helton, W. S. (2011). Animal expertise: Evidence of phase transitions by utilizing running estimates of performance variability. *Ecological Psychology, 23*(2), 59–75.

Hofmann, D. A., Griffin, M. A., & Gavin, M. (2000). The application of hierarchical linear modeling to management research. In K. Klein & S. W. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 467–511). San Francisco, CA: Jossey-Bass.

Hoogervorst, J. (2004). Enterprise architecture: Enabling integration, agility and change. *International Journal of Cooperative Information Systems, 13*(3), 213–233.

ICAO. 2013. Global Air Transport Outlook to 2030 and Trends to 2040. No. Cir 333, AT/190. Montreal, Canada.

Janczura, J., & Weron, R. (2012). Black swans or dragon-kings? A simple test for deviations from the power law. *The European Physical Journal Special Topics, 205*(1), 79–93.

Jenatabadi, H. S., & Ismail, N. A. (2007). The determination of load factors in the airline industry. *International Review of Business Research Papers, 3*(4), 125–133.

Jones, S., Kirchsteiger, C., & Bjerke, W. (1999). The importance of near miss reporting to further improve safety performance. *Journal of Loss Prevention in the Process Industries, 12*(1), 59–67.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity, 47*(4), 2025–2047.

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, BUGS and Stan*. Amsterdam, the Netherlands: Elsevier.

Laney, D. B. (2002). Improved control charts for attributes. *Quality Engineering, 14*, 531–537.

Larcos, G., Collins, L. T., Georgiou, A., & Westbrook, J. I. (2014). Maladministrations in nuclear medicine: Revelations from the Australian radiation incident register. *Medical Journal of Australia, 200*, 37–40.

Larcos, G., Collins, L. T., Georgiou, A., & Westbrook, J. I. (2015). Nuclear medicine incident reporting in Australia: Control charts and notification rates inform quality improvement. *Internal Medicine Journal, 45*(6), 609–617.

Lenne, M. G., Salmon, P. M., Liu, C. C., & Trotter, M. (2012). A systems approach to accident causation in mining: An application of the HFACS method. *Accident Analysis & Prevention, 48*, 111–117.

Levinson, D. (2012). *Hospital incident reporting systems do not capture most patient harm*. Office of Inspector General, Department of Health and Human Services.

Lindberg, A. K., Hansson, S. O., & Rollenhagen, C. (2010). Learning from accidents–what more do we need to know?. *Safety Science, 48*(6), 714–721.

Loft, S., Sanderson, P., Neal, A., & Mooij, M. (2007). Modeling and predicting mental workload in en route air traffic control: Critical review and broader implications. *Human Factors, 49*, 376–399.

Ma, J., Pedigo, M., Blackwell, L., Gildea, K., Holcomb, K., Hackworth, C., & Hiles, J. J. (2011). *The line operations safety audit program: Transitioning from flight operations to maintenance and ramp operations* (No. DOT/FAA/AM-11-15). Washington, DC: Federal Aviation Administration.

Madsen, P. M. (2013). Perils and profits a reexamination of the link between profitability and safety in US Aviation. *Journal of Management, 39*(3), 763–791.

Marais, K. B., & Robichaud, M. R (2012). Analysis of trends in aviation maintenance risk: An empirical approach. *Reliability Engineering and System Safety, 106*, 104–118.

Montgomery, D. (2012). *Introduction to statistical quality control*. New York: Wiley.

Neal, A., Hannah, S., Sanderson, S., Bolland, S., Mooij, M, & Murphy, S. (2014). Development and validation of a multilevel model for predicting workload under routine and nonroutine conditions in an air traffic management center. *Human Factors, 56*, 287–305.

O'Connor, P., Buttrey, S. E., O'Dea, A., & Kennedy, Q. (2011). Identifying and addressing the limitations of safety climate surveys. *Journal of safety research, 42*(4), 259–265.

O'Hare, D., Wiggins, M., Batt, R., & Morrison, D. (1994). Cognitive failure analysis for aircraft accident investigation. *Ergonomics, 37*(11), 1855–1869.

Ohrn, A., Elfstrom, J., Liedgren, C., & Rutberg, H. (2011). Reporting of sentinel events in Swedish hospitals: A comparison of severe adverse events reported by patients and providers. *Joint Commission Journal on Quality & Patient Safety, 37*(11), 495–501.

Oster, C. V., Strong, J. S., & Zorn, C. K. (2013). Analyzing aviation safety: Problems, challenges, opportunities. *Research in Transportation Economics, 43*(1), 148–164.

Ramzy, A. (February 12, 2015). Many pilots fail safety test at TransAsia. *New York Times*, p. A8.

Reddy, S. B., & Reddy, R. (2002). Competitive agility and the challenge of legacy information systems. *Industrial Management & Data Systems, 102*(1), 5–16.

Revelle, W., Humphreys, M. S., Simon, L., & Gilliland, K. (1980). The interactive effect of personality, time of day, and caffeine: A test of the arousal model. *Journal of Experimental Psychology: General, 109*, 1–31.

Roth, W. M. (2015). Cultural practices and cognition in debriefing the case of aviation. *Journal of Cognitive Engineering and Decision Making, 9*, 263–278.

Sachs, M. K., Yoder, M. R., Turcotte, D. L., Rundle, J. B., & Malamud, B. D. (2012). Black swans, power laws, and dragon-kings: Earthquakes, volcanic eruptions, landslides, wildfires, floods, and SOC models. *The European Physical Journal-Special Topics, 205*(1), 167–182.

Shojania, K. (2008). The frustrating case of incident-reporting systems. *Quality & Safety in Health Care, 17*(6), 400–402.

Shorrock, S. T. (2005). Errors of memory in air traffic control. *Safety Science, 43*, 571–588.

Shorrock, S. T. (2007). Errors of perception in air traffic control. *Safety Science, 45*, 890–904.
Shorrock, S. T., & Kirwan, B. (2002). Development and application of a human error identification tool for air traffic control. *Applied Ergonomics, 33*, 319–336.
Simons, G. M. (2013). *Comet: The world's first jet airliner.* Barnsley, UK: Pen and Sword.
Simoudis, E. (1996). Reality check for data mining. *IEEE Intelligent Systems, 11*(5), 26–33.
Sornette, D., & Ouillon, G. (2012). Dragon-kings: Mechanisms, statistical methods and empirical evidence. *The European Physical Journal Special Topics, 205*(1), 1–26.
Stanhope, N., Crowley-Murphy, M., Vincent, C., O'Connor, A. M., & Taylor-Adams, S. E. (1999). An evaluation of adverse incident reporting. *Journal of Evaluation in Clinical Practice, 5*(1), 5–12.
Sull, D. (1999). Easyjet's $500 million gamble. *European Management Journal, 17*(1), 20–32.
Taleb, N. N., (2007). *The black swan: The impact of the highly improbable*. London, UK: Penguin.
Tariq, A., Georgiou, A., & Westbrook, J. (2013). Medication errors in residential aged care facilities: A distributed cognition analysis of the information exchange process. *International Journal of Medical Informatics, 82*(5), 299–312.
Thomas, M. J., & Petrilli, R. M. (2006). Crew familiarity: Operational experience, non-technical performance, and error management. *Aviation, Space, and Environmental Medicine, 77*(1), 41–45.
Travaglia, J. F., Nugus, P. I., Greenfield, D., Westbrook, J. I., & Braithwaite, J. (2011). Visualising differences in professionals' perspectives on quality and safety. *BMJ Quality & Safety, 21*, 778–783.
Tretheway, M. W., & Markhvida, K. (2014). The aviation value chain: Economic returns and policy issues. *Journal of Air Transport Management, 41*, 3–16.
Walker, G., & Strathie, A. (2016). Big data and ergonomics methods: A new paradigm for tackling strategic transport safety risks. *Applied Ergonomics, 53*, 298–311.
Wang, H., & Gao, J. (2013). Bayesian network assessment method for civil aviation safety based on flight delays. *Mathematical Problems in Engineering, 2013,* 1–12.
Westbrook, J. I., Gosling, A. S., & Coiera, E. (2004). Do clinicians use online evidence to support patient care? A study of 55,000 clinicians. *Journal of the American Medical Informatics Association, 11*(2), 113–120.
Westbrook, J. I., Li, L., Lehnbom, E., Braithwaite, B. M. J., Burke, R., Conn, C., & Day, R. (2015). What are incident reports telling us? A comparative study at two Australian hospitals of medication errors identified at audit, detected by staff and reported to an incident system. *International Journal of Quality in Health Care, 27*(1), 1–9.
Wiegmann, D. A., & Shappell, S. A. (2001). Human error analysis of commercial aviation accidents: Application of the human factors analysis and classification system (HFACS). *Aviation, Space, and Environmental Medicine, 72*(11), 1006–1016.
Wiggins, M. W., Hunter, D. R., O'Hare, D., & Martinussen, M. (2012). Characteristics of pilots who report deliberate versus inadvertent visual flight into instrument meteorological conditions. *Safety Science, 50*(3), 472–477.
Wiggins, M. W., & Stevens, C. (1999). *Aviation social science: Research methods in practice.* Aldershot, UK: Ashgate Publishing.
Wilke, S., Majumdar, A., & Ochieng, W. Y. (2014). A framework for assessing the quality of aviation safety databases. *Safety Science, 63*, 133–145.
Withey, P. A. (1997). Fatigue failure of the de Havilland comet I. *Engineering Failure Analysis, 4*(2), 147–154.